

Report2

Load the R packages

```
library(knitr)
library(readr)
library(tidyverse)
library(dplyr)
library(tidyr)
library(stringr)
```

Load the data

```
data <- read_csv("survey.csv")

data_types <- data %>%
  dplyr::summarise_all(class) %>%
  tidyr::gather(variable, class)

participants_per_university <- data %>%
  group_by(University) %>%
  summarize(total_participants = n())
```

Part 1

They were **30** survey participants.

They survey is made up of the following data types: **numeric, character**.

The top 3 universities with the most survey participants were

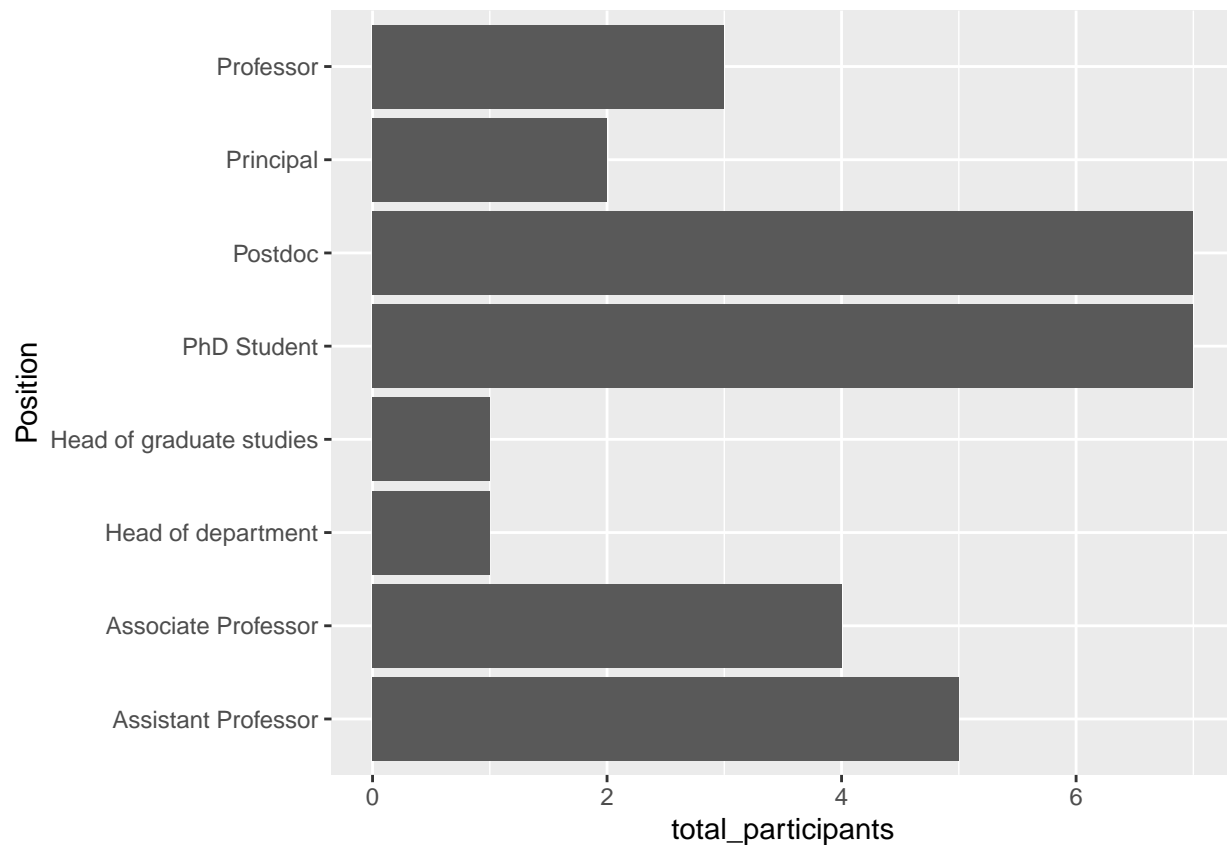
University	total_participants
Delft University of Technology	9
Eindhoven University of Technology	7
Erasmus University Rotterdam	5

Part 2

A horizontal bar graph of the number of survey participants for the different positions given in the column *Position*:

```
participants_per_position <- data %>%
  group_by(Position) %>%
  summarize(total_participants = n())

ggplot(participants_per_position, aes(x=Position, y=total_participants)) + geom_bar(position="dodge", stat="sum")
```



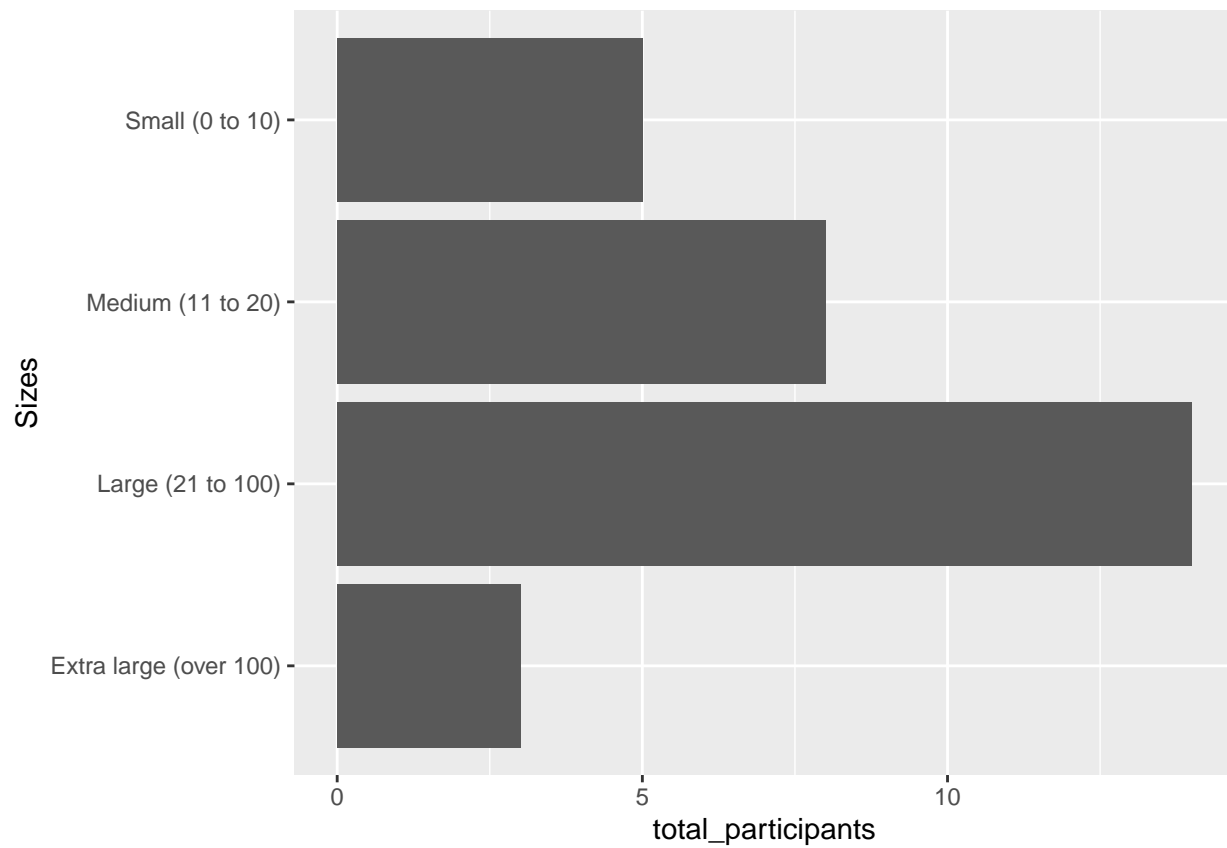
Part 3

```
# define labels
label_s <- "Small (0 to 10)"
label_m <- "Medium (11 to 20)"
label_l <- "Large (21 to 100)"
label_xl <- "Extra large (over 100)"

# Recreate the Sizes column
for (ind in seq(1,nrow(data))) {
  if (data$'Number of employees'[ind] <= 10) {
    data$'Sizes'[ind] <- label_s
  } else if (data$'Number of employees'[ind] <= 20) {
    data$'Sizes'[ind] <- label_m
  } else if (data$'Number of employees'[ind] <= 100) {
    data$'Sizes'[ind] <- label_l
  } else {
    data$'Sizes'[ind] <- label_xl
  }
}

participants_per_sizes <- data %>%
  group_by(Sizes) %>%
  summarize(total_participants = n())
```

```
# Plot
ggplot(participants_per_sizes, aes(x=Sizes, y=total_participants)) + geom_bar(position="dodge", stat="id
```



Part 4

```
# Mutate data with new aggregated values
data_share <- data %>%
  rename(
    n_total    = 'Number of employees',
    n_phds     = 'Number of PhD students',
    n_postdocs = 'Number of postdocs',
    n_profs    = 'Number of professors'
  ) %>%
  mutate(
    share_phds    = 100*n_phds/n_total,
    share_postdocs = 100*n_postdocs/n_total,
    share_profs   = 100*n_profs/n_total,
    share_total   = 100
  ) %>%
  select(X1, share_phds, share_postdocs, share_profs) %>%
  as_tibble()

data_gathered <- data_share %>%
```

```
gather(key = "Type",
       value = "Percentage",
       share_phds,
       share_postdocs,
       share_profs)

data_gathered <- data_gathered %>%
  rename(Research_Group_ID = 'X1')

ggplot(data=data_gathered, aes(x=Research_Group_ID, y = Percentage, fill=Type)) +
  geom_bar(stat="identity") + coord_flip()
```

